3M

УДК 32.973.202:07.681

О.О. Марченко[1], О.М. Марченко-Бабіч[2]
[1]Київський національний університет імені Тараса Шевченка,
факультет кібернетики, Україна
Україна, 03680, м. Київ, просп. Глушкова, 4-д
[2]Військовий інститут Київського національного університету ім. Тараса Шевченка,
м. Київ, Україна
Україна, 03680, просп. Глушкова, 2

Формування лінгвістичного забезпечення для автоматизованого пошуку і відбору текстів на ресурсах новин та в соціальних мережах

O.O. Marchenko[1], O.M. Marchenko-Babich[2]
[1]Taras Shevchenko National University of Kyiv, Faculty of Cybernetics, Ukraine
Ukraine, 03680, c. Kyiv, Glushkova Ave., 4-d
[2] The Military Institute of Taras Shevchenko National University of Kyiv, Ukraine
Ukraine, 03680, c. Kyiv, Glushkova Ave., 2

Development of Linguistic Software for Automated Search and Selection of Texts on News Resources and in Social Networks

А.А. Марченко[1], О.Н. Марченко-Бабич[2]
[1]Киевский национальный университет имени Тараса Шевченко, Украина
Украина, 03680, г. Киев, просп. Глушкова, 4-д
[2]Военный институт Киевского национального университета им. Тараса Шевченко,
г. Киев, Украина
Украина, 03680, просп. Глушкова, 2

Формирование лингвистического обеспечения для автоматизированного поиска и отбора текстов на новостных ресурсах и в социальных сетях

У статті розглядаються нові принципи пошуку і відбору текстів когнітивних технологій у мережі Інтернет. Описано компоненти бази знань системи автоматизованого моніторингу. Запропоновано підходи до побудови методів семантичного пошуку текстів з елементами сугестії.
Ключові слова: обробка природної мови, пошук текстів, сугестія.

The article concerns the new principles of search and selection of cognitive technologies texts in the Internet. The components of the knowledge base for the automated monitoring system are described. The approaches to developing of techniques for semantic search of texts with suggestion elements are proposed.
Keywords: natural language processing, text searching, suggestion.

В статье рассматриваются новые принципы поиска и отбора текстов когнитивных технологий в сети Интернет. Описаны компоненты базы знаний системы автоматизированного мониторинга. Предложены подходы к построению методов семантического поиска текстов с элементами суггестии.
Ключевые слова: обработка естественного языка, поиск текстов, суггестия.

Останнім часом серед найбільш актуальних напрямів у розробці інформаційного забезпечення виділився такий помітний сегмент IT-досліджень, як створення технологій шостого покоління, які в англомовній термінології мають абревіатуру NBIC (в українській термінології використовують абревіатуру НБІК-технології), відповідно до перших букв найменувань технологій: N або Н – нано, B або Б – біо, I – інфо, C або К – когно. Даний підхід має інтегрувати в собі риси та характеристики всіх вищенаведених технологій для побудови та аналізу об'ємної багатовимірної моделі інформаційних об'єктів. Побудована модель за допомогою багаторівневого представлення різної за походженням та модальністю інформації про об'єкт виводить якість аналізу та інтелектуальної обробки текстів на суттєво вищий рівень з огляду на наявність різнопланових точок зору на інформаційні об'єкти вхідного тексту.

Враховуючи, що останнім часом обсяги інформаційних потоків подвоюються менш ніж за 5 років [1], то стратегічно важливим є не стільки володіння інформацією, скільки вміння швидше за інших її обробити, систематизувати й отримати нові актуальні знання. На вирішення цієї задачі зорієнтовані когнітивні технології, які, у свою чергу, можуть бути побудовані на основі формалізації когнітивних здібностей людини (лат. cognitio – пізнання, пізнавання, пізнавальні функції). Когнітивні технології часто реалізуються через засоби масової інформації (ЗМІ), найбільш оперативним з яких є Інтернет. Залежно від спрямованості пошуку, потрібна інформація міститься як на сайтах новин, так і різноманітних соціальних мережах, форумах, блогах, інших площадках Інтернет-спілкування. Їх аудиторія становить мережне співтовариство та може бути як їх активним учасником, так і об'єктом їх застосування.

Оскільки визначальним чинником у використанні когнітивних технологій є швидкість опрацювання інформації з її подальшим оптимальним використанням, потрібно удосконалення технологій пошуку та опрацювання інформаційних повідомлень ЗМІ за допомогою програмних засобів автоматизованого моніторингу. Для цього доцільним є винайдення нових підходів та вдосконалення вже існуючих у лінгвістичному забезпеченні цих засобів, залежно від інформаційних потреб користувача (аналітика служби моніторингу).

Тому **метою і основним змістом статті** є розробка нових принципів пошуку та відбору інформації серед текстів когнітивних технологій у мережі Інтернет та нових підходів до формування лінгвістичного та програмного забезпечення системи автоматизованого моніторингу.

При формуванні лінгвістичного забезпечення для пошуку зазначених текстів необхідно врахувати їх особливості: як лексичні, так і особливості мовних конструкцій. Важливою властивістю текстів когнітивних технологій, присутніх в Інтернет, є *передавання інформації за допомогою частково неусвідомлюваного, направленого сигналу на вербальному чи невербальному рівнях – так звана сугестія*. Це форма міжособистісного та міжгрупового спілкування, яка відрізняється від переконання зниженим рівнем критичності та потреби у верифікації інформації.

Застосування в Інтернет когнітивних технологій з елементами сугестії націлено на масовий результат. Їхнім об'єктом найчастіше виступає певне мережне співтовариство: соціальні мережі, форуми, блоги, інші площадки Інтернет-спілкування.

Європейська дослідницька компанія InsitesConsulting.eu підрахувала, що різними **соціальними мережами** в усьому світі зараз користуються більш ніж 1 млрд людей [2]. Останнім часом до соціальних мереж приєдналось більш ніж 70% усіх Інтернет-користувачів. Хвиля «facebook»-революцій дійсно продемонструвала здатність Інтернету грати

провідну роль серед ЗМІ, оскільки це наймасовіший, найдешевший та найважче контрольований державою засіб масової інформації. Події «facebook»-революцій, що відбулись у низці країн в останні роки, довели, що сугестивний вплив на мережні співтовариства може спонукати їх представників до дій у реальному світі.

За дотримання сугестивності на мовному рівні відповідає **сугестивна лінгвістика** – міждисциплінарна наука на стику філології та психології. Формою втілення сугестивності у мові є дискурс. Він може бути вербальним і невербальним (жести, міміка тощо). Особливостями такого дискурсу в Інтернет-новинах та на сайтах соцмереж є:

– конкретність та образність ключових слів у дискурсах офіційних і неофіційних Інтернет-ресурсів і посиланнях пошукових систем. В інтернетних посиланнях автори новин намагаються використати ключові слова, які навіть у випадку непрочитання самого тексту програмували б читача у потрібному напрямі;

– емоційне перенасичення тексту: велика кількість яскравих прикметників, порівнянь, метафор та інших образних засобів, що підмінюють фактичний матеріал;

– використання риторичних запитань, що підштовхують читача до потрібних відповідей;

– приховування джерел інформації (з посиланням на «деяких експертів» тощо);

– вживання наказових конструкцій, що найбільш дієво для посттоталітарного співтовариства;

– використання лінгвістичних структур єдності, довіри;

– експлуатація ідеї «кола своїх», навмисне включення до нього мережного ресурсу;

– використання евфемізмів, що залучає підсвідомість споживача інформації та формує необхідний маніпулятору образ;

– активне звертання до антропоцентричних словотвірних моделей: уведення в текст новоутворень, що називають осіб;

– представлення слова як фізичного тіла, яке може стискатись, розширюватись та зливатись з іншими словами;

– візуальне підкріплення змісту переданої інформації також сприяє підвищенню сили навіювання (певним чином підібраними фотографіями, малюнками, смайлами тощо замінюють в Інтернет-комунікації міміку, жестикуляцію), оскільки немовні моменти спілкування менше піддаються осмисленому контролю: «ні з чого» виникає певне емоційне ставлення.

З огляду на це формування бази знань системи автоматизованого моніторингу повинно враховувати як лексичні одиниці, так і мовні конструкції, властиві сугестивному дискурсу. Воно також повинно відображати загальні тенденції розвитку у сфері, якої стосуються повідомлення зазначеної специфіки. Цю базу слід періодично оновлювати та настроювати згідно з поточним інформаційним контентом.

До бази знань даної системи слід включити такі складові:

– визначений профіль пошуку (розділи новин в електронних ЗМІ, певні сайти, що є місцем спілкування Інтернет-спільнот);

– тематику, визначену напрямом діяльності;

– емоційну забарвленість текстів повідомлень;

– мову повідомлення, визначену завданням пошуку.

Одним із засобів, здатних встановити емоційну забарвленість тексту, є технологія Sentiment analysis, що дозволяє розподілити повідомлення за характером на позитивні та негативні згідно з оціночними судженнями їх авторів про предмет обговорення. Завдяки фільтрам на інформацію певного характеру, наприклад, негативну, є можливість відбирати тексти певної спрямованості згідно з завданнями пошуку.

При здійсненні моніторингу необхідно враховувати пошуковий профіль користувача (із врахуванням особливостей спектра його інтересів), а також те, щоб можна було задавати пошук не лише запитами, але й прикладами еталонних документів за їх «образом і подобою» за змістом і за семантикою. Новизна підходу полягає у тому, що документ шукається не за принципом співпадіння ключових слів, а за принципом відповідності семантичних структур знайденого документа запиту користувача. Саме завдяки цьому вдається ефективно долати проблеми негативного впливу полісемії багатозначних слів та словосполучень на точність смислового аналізу текстів [3].

Також важливим є те, що пропонується запровадження алгоритмів семантичного пошуку, які допомагають поширювати інформаційно-пошукові запити за допомогою синонімів, семантично-близьких понять (термів), які містяться в семантичній базі знань системи [4]. Це дасть змогу формалізувати процес складання ефективного пошукового запиту, побудувати синонімічний ряд для кожного зі слів та вкласти до пошукової системи усі необхідні дані. Таким чином можна знайти такий текст, який не містить жодного ключового слова з запиту і при цьому повністю за змістом та семантикою відповідає даному запиту.

Для знаходження повідомлень за вказаним напрямом із текстом, що «підозрюється» на наявність ознак сугестивності, проводиться лінгвістичний аналіз, складовими якого є лексико-морфологічний, синтаксичний, семантичний аналіз для отримання певної семантичної структури, яку можна проаналізувати з точки зору впливу на цільову аудиторію. Згідно з синтаксичною структурою текстів будуються семантичні графи та проводиться психолінгвістичний аналіз їх компонентів.

Повідомлення, якими обмінюються в Інтернет-спільноті, часто представлені у вигляді коротких текстів, (наприклад, «твітів»), які не піддаються стандартним алгоритмам. Для відстеження даних повідомлень необхідно використання алгоритмів, спеціально пристосованих для обробки таких текстів. Тому з урахуванням нових особливостей Інтернет-контенту з'являється все більше спеціалізованих пошукових систем, які використовують для пошуку на сайтах з конкретної тематики. Згідно з останніми науковими дослідженнями, для побудови бази знань для вирішення зазначених завдань доцільно використовувати як тексти довільної форми, так і напівструктуровані джерела інформації (таблиці, списки, сайти регулярної структури). Також слід приділити увагу системам безперервного навчання, наприклад, такій, що реалізована у проекті NELL [5] та ітераційно виконує дві задачі: задачу читання і задачу навчання. Під задачею читання розуміється отримання системою нових фактів з неструктурованих або напівструктурованих джерел (текстів). Задача навчання – на отриманих фактах сформувати нові патерни для більш ефективного «читання» системою текстових масивів мережі Інтернет [6].

Автоматичний розподіл відібраних повідомлень доцільно здійснювати залежно від особливостей висвітлення у них об'єкта пошуку. Для цього у процесі семантичного аналізу повідомлень з елементами сугестивних технологій пропонується використання підходів, що застосовуються у **системах семантичного моніторингу** [7]. В даних системах використовуються контекстний асоціативно-семантичний аналіз для обробки текстових потоків і корпусів з блоком якісного оцінювання лінгвістичних фокусних об'єктів. Він дозволяє обчислювати якісні характеристики й параметри будь-якого заданого лінгвістичного об'єкта в корпусах текстів і текстових потоках, відстежуючи динаміку змін та визначаючи основні тенденції оцінювання фокусного об'єкта. Після подачі на вхід системи імені заданого об'єкта, вона формує семантичний фокус-образ у мережі онтології, обчислюючи якісні характеристики і параметри заданого об'єкта

в тексті. Важливим етапом створення системи семантичного моніторингу є формування лінгвістичної шкали для якісних оціночних концептів онтології. Перший підхід визначення чисельно-порядкових значень концептів виконується за допомогою асоціативно-контекстних алгоритмів, які шукають відстані в мережі онтології між поточним концептом і концептом-максимумом (мінімумом) даної шкали. Другий підхід, задіяний при розробці лінгвістичної шкали, використовує частотні алгоритми, що визначають частоту спільної появи пар слів у глобальних корпусах текстів, встановлюючи таким чином близькість їх семантичних значень (із врахуванням винятків серед сполучень певних груп слів). Такий контекстний асоціативно-семантичний аналіз дозволяє гнучко варіювати значення якісних оціночних концептів, залежно від локально-глобального контексту, що дає можливість враховувати складні з точки зору ординарної семантики випадки застосування лексики.

Використання цього підходу забезпечує обчислення якісних характеристик і параметрів тексту з відстеженням динаміки змін та визначенням основних тенденцій оцінювання об’єкта вивчення.

Висновки

Запровадження програмних засобів пошуку та відбору текстових повідомлень когнітивних технологій у мережі Інтернет передбачає розробку нових підходів до створення лінгвістичного забезпечення. Важливою властивістю текстів когнітивних технологій є сугестія, що враховує інформаційні потреби користувачів, тому лінгвістичне забезпечення повинно включати особливості сугестивного дискурсу. Це як певні лексичні одиниці, так і мовні конструкції, які слід брати до уваги при формуванні бази знань системи автоматизованого моніторингу. Цю базу слід періодично оновлювати та настроювати згідно з поточним інформаційним контентом.

Отже, лінгвістичне забезпечення програмних засобів пошуку та відбору текстових повідомлень когнітивних технологій повинно включати:

– базу знань із врахуванням профілю пошуку, об’єктів пріоритетного вивчення Інтернет-повідомлень та їх особливостей;

– алгоритми як для обробки неструктурованих даних (звичайних текстів новин тощо), так і для напівструктурованих даних (таблиць, списків, сайтів регулярної структури) та коротких повідомлень.

Основні етапи обробки текстових повідомлень з елементами сугестії повинні включати:

– формалізацію повідомлень за напрямами пошуку, що передбачає побудову семантичних графів згідно із синтаксичною структурою речень у текстах повідомлень та подальшим психолінгвістичним аналізом компонентів графу;

– автоматичний розподіл повідомлень, відібраних з мережі Інтернет програмними засобами, з урахуванням актуальності повідомлення та характеристик джерел, які їх поширюють, а також особливостей висвітлення об’єктів, що становлять інтерес.

При здійсненні моніторингу слід враховувати пошуковий профіль користувача, особливості спектра його інтересів, можливість завдання пошуку не лише запитами, але й прикладами еталонних документів за їх «образом і подобою». Пошук документа доцільно проводити не за принципом простого співпадіння ключових слів, а за принципом відповідності семантичних структур знайденого документа запиту користувача. Запровадження алгоритмів семантичного аналізу дозволить поширювати інформаційно-

пошукові запити за допомогою синонімів, семантично-близьких понять (термів), які містяться в семантичній базі знань системи. Застосування підходів напряму Sentiment Analysis [8] в системах моніторингу та пошуку текстів когнітивних технологій дає можливість врахування багатьох нюансів та деталей емоційного забарвлення текстових повідомлень, що є дуже затребуваним з огляду на специфіку основних цілей даної системи.

Література

1. [Електронний ресурс]. – Режим доступу: http://nvo.ng.ru/concepts/2011-12-02/6_nanobioinfo.html
2. Сугестивні технології маніпулятивного впливу : [навчальний посібник] / [В.М. Петрик, М.М. Присяжнюк, Л.Ф. Компанцева та інш.] ; за заг. ред. Є.Д. Скулиша − К. : Науково-видавничий відділ НА СБ України, 2010. − 248 с.
3. Марченко О.О. Моделювання семантичного контексту при аналізі текстів на природній мові / О.О. Марченко // Вісник Київського університету. Сер. фіз.-мат. науки. – 2006. – № 3. – С. 230-234.
4. Анісімов А.В. UWN: Універсальна онтологічна база знань укранської мови / А.В. Анісімов, О.О. Марченко, А.О. Никоненко // Проблеми програмування. – 2012. – № 2 – 3. – С. 348-355.
5. [Електронний ресурс]. – Режим доступу: http://www.cmu.edu/homepage/computing/2010/fall/nell-computer-that-learns.shtml.
6. Глибовець А.М. Алгоритми обробки текстів вільної форми для отримання фактів і зв'язків між ними / [Глибовець А.М., Марченко О.О., Циганок Д.В., Бабіч О.М.] // Наукові записки НаУКМА. Комп'ютерні науки. – 2012. – Т. 138. – С. 35-38.
7. Марченко А.А. Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта / А.А. Марченко, А.А. Никоненко // Искусственный интеллект. − 2008. − Вип. 3.− С. 808-813.
8. Bo Pang. Opinion mining and sentiment analysis / Bo Pang and Lillian Lee // Foundations and Trends in Information Retrieval. – 2008. – Vol. 2, № 1 – 2. – P. 1-135.

Literatura

1. http://nvo.ng.ru/concepts/2011-12-02/6_nanobioinfo.html
2. Suggestion technologies of manipulation influence: tutorial / [V.M. Petric, M.M. Prysiagnyuk, L.F. Compantseva, E.D. Skulysh, O.D.Boyko, V.V. Ostrouchov]; under the gen. editorship of E.D. Skulysh. − К. : research and publishing department of NA SBU, 2010. − 248 p.
3. Marchenko O.O. Modeling of semantic context in the analysis of natural language texts. Bulletin of Taras Shevchenko National University of Kyiv. Ser. phis.-math. sci. – 2006. – № 3. – P. 230-234.
4. Anisimov A.V., Marchenko O.O., Nykonenko A.O. UWN: Universal ontological knowledge base of Ukrainian language. Problems of programming. – 2012. – № 2 – 3. – P. 348-355.
5. http://www.cmu.edu/homepage/computing/2010/fall/nell-computer-that-learns.shtml
6. Glibovets A.M. Algorithms for processing free-form text for extracting facts and links between them / Glibovets A.M, Marchenko O.O., Tsyganok D.V., Babich O.M. // Proceedings NaUKMA. Computer science.– 2012. – Vol. 138. – P. 35-38.
7. Marchenko O.O., Nykonenko A.O. Context semantic analysis of text. System for text monitoring and quality evaluation of focal object. // Artificial intelligence. − 2008. − № 3.− P. 808-813.
8. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – Vol. 2, № 1 – 2. – 2008. – P. 1-135.

RESUME

O.O. Marchenko, O.M. Marchenko-Babich

Development of Linguistic Software for Automated Search and Selection of Texts on News Resources and in Social Networks

New approaches to software development, and in particular the linguistic component, are needed for search and selection of text messages of special type as news sites and social

networks. These texts are the part of cognitive technologies, and the suggestion is an important element of them, therefore, the linguistic support should incorporate models that concern features of suggestive discourse. The main elements of the knowledge base for software search and selection of texts from news sites and social networks are:

– knowledge base, taking into account the profile of the search, priority research objects of Internet messages and their features;

– algorithms for data processing, both unstructured and semi-structured, as well as for short messages processing.

The main stages of texts processing with the elements of suggestion should include:

– formalization of the posts in the search directions, which include generation of semantic graphs according to the syntactic structure of sentences in the message, followed by psycholinguistic analysis of the graph components;

– automatic distribution of messages captured in the Internet with software taking into account their actuality and the characteristics of their sources and features of the objects description.

The monitoring should consider search user profile features, his interests spectrum, the ability to provide search queries not only by keywords coincidence, but also by examples of reference documents. Document search should be carried out not only by the principle of simple matching of keywords, but also by the principle of matching the semantic structures of the document to the user's query. Using semantic analysis algorithms will provide an opportunity to expand queries with synonyms, semantically-close concepts (terms) contained in the semantic-based systems. Applying techniques of Sentiment Analysis in monitoring and searching of cognitive technologies texts will take into account the many nuances and details of the emotional modality of texts from news sites and social networks.

Стаття надійшла до редакції 12.04.2013.